CAMBRIDGE
UNIVERSITY PRESS

Article

Effects of Genetic Relatedness of Kin Pairs on Univariate ACE Model Performance

Xuanyu Lyu[1,2,3] ![ORCID] and S. Mason Garrison[1]

[1]Department of Psychology, Wake Forest University, Winston Salem, North Carolina, USA, [2]Institute for Behavioral Genetics, University of Colorado at Boulder, Boulder, Colorado, USA and [3]Department of Psychology & Neuroscience, University of Colorado at Boulder, Boulder, Colorado, USA

Abstract

The current study explored the impact of genetic relatedness differences (ΔH) and sample size on the performance of nonclassical ACE models, with a focus on same-sex and opposite-sex twin groups. The ACE model is a statistical model that posits that additive genetic factors (A), common environmental factors (C), and specific (or nonshared) environmental factors plus measurement error (E) account for individual differences in a phenotype. By extending Visscher's (2004) least squares paradigm and conducting simulations, we illustrated how genetic relatedness of same-sex twins (H_{SS}) influences the statistical power of additive genetic estimates (A), AIC-based model performance, and the frequency of negative estimates. We found that larger H_{SS} and increased sample sizes were positively associated with increased power to detect additive genetic components and improved model performance, and reduction of negative estimates. We also found that the common solution of fixing the common environment correlation for sex-limited effects to .95 caused slightly worse model performance under most circumstances. Further, negative estimates were shown to be possible and were not always indicative of a failed model, but rather, they sometimes pointed to low power or model misspecification. Researchers using kin pairs with ΔH less than .5 should carefully consider performance implications and conduct comprehensive power analyses. Our findings provide valuable insights and practical guidelines for those working with nontwin kin pairs or situations where zygosity is unavailable, as well as areas for future research.

Statistical power, an integral part of the research process, underpins the design and evaluation of empirical research. Beyond just good science, it is a near universal expectation from granting agencies (Chow et al., 2017; Descôteaux, 2007). These agencies typically expect researchers to determine adequate sample sizes through a priori power calculations (Cohen, 1988; Jackson et al., 2009; Maxwell et al., 2008). Post hoc calculations are equally important — they facilitate evaluating the conclusions drawn from any given study (Levine & Ensom, 2001). Within the field of behavior genetics, particularly in the context of ACE models, power refers to the probability of correctly rejecting the null hypothesis — that genetic or common environmental effects have no impact on the outcome traits (Verhulst, 2017; Visscher, 2004). The ACE model is a statistical model that posits that additive genetic factors (A), common environmental factors (C), and specific (or nonshared) environmental factors plus measurement error (E) account for individual differences in a phenotype. Previous studies have thoroughly discussed how sample size, variance components, and the ratio of twin types impact the power of parameter estimation via mathematical derivations (e.g., Visscher, 2004; Visscher et al., 2008), computer simulations

(e.g., Verhulst, 2017), or a combination of the two (Martin et al., 1978; Sham et al., 2020). Notably, these studies all use monozygotic (MZ) and dizygotic (DZ) twins. However, an ACE model is not exclusively identifiable with MZ and DZ twins. In fact, any two groups of kin pairs with different genetic-relatedness parameters (H) are mathematically sufficient to fit an ACE model (Hunter et al., 2021).

Most simulation research has focused on classical twin designs, which set the genetic relatedness parameter (H) at 1.0 for MZ twins and 0.5 for DZ twins. However, alternate family designs employing different H parameters do exist. One such example is the same-sex (SS) and opposite-sex (OS) DZ twin pair design, often called the SS-OS design. This design becomes particularly useful when zygosity is unavailable as researchers can distinguish OS DZ twins from SS twins based on birth date and biological sex. The remaining SS twin pairs are a mixture of MZ and SS DZ twins. Historically, this design was a staple in earlier twin studies such as the Scottish Mental Surveys conducted in 1932 and 1947 (Deary et al., 2004), before the widespread use of genotyping. Despite technological advancements, this design remains relevant, particularly when genotyping is not feasible. For example, a series of studies by Figlio and colleagues (Figlio, Guryan et al., 2014; Figlio, Freese et al., 2017) used the SS-OS twin design on administrative data to analyze all twins born in Florida from 1994 to 2004. The authors relied on

these records to increase the representation of twins from disadvantaged backgrounds, thereby mitigating selection effects commonly found in twin studies (Hagenbeek et al., 2023; Holden et al., 2022). In doing so, they ensured a more representative sample, but they had to forgo determining zygosity — a design trade-off the authors argued was worthwhile (Figlio et al., 2017).

These design considerations are not unique to Figlio and colleagues (Figlio, Guryan et al., 2014; Figlio, Freese et al., 2017), but reflect a widespread limitation, as most surveys lack zygosity data. Yet, many large-scale social surveys[1] collect family data without being specifically tailored for twin studies. These surveys usually focus on social, economic, educational, geographical, and political topics (e.g., China Family Panel Study, by Xie & Hu, 2014; The National Longitudinal Survey of Youth, by Rodgers et al., 2016), and employ household sampling methods for efficiency (Parsaeian et al., 2021; United Nations, 2008). By deploying the SS-OS design on these public datasets, we enable these datasets to yield not just individual-level information but also rich genetic insights from twin studies, adding depth to our analyses. Compared to twin registries and genomewide association studies (GWASs), these public datasets often cover a wider range of research topics and contain more diverse populations (Hagenbeek et al., 2023; Holden et al., 2022), allowing us to move beyond typical WEIRD (Western, Educated, Industrialized, Rich, and Democratic) samples (Henrich et al., 2010; Popejoy & Fullerton, 2016; see Holden et al., 2022; Milhollen et al., 2022 for additional discussion of these samples for behavior geneticists). Compared to individual-level analysis, another notable advantage of the SS-OS design, such as the one used by Figlio and colleagues (Figlio, Guryan et al., 2014; Figlio, Freese et al., 2017), is its capacity to meet the equal environments assumption, without exclusively relying on MZ versus DZ twins.[2]

The genetic relatedness patterns for twins ($H_{MZ} = 1.0$ and $H_{DZ} = .5$) are a byproduct of their development (Beck et al., 2021). DZ twins, on average, share 50% of their segregating genes, a percentage that arises from the random segregation of each chromosome pair within the gametes. MZ twins, conversely, share 100% of their genes as they originate from the same zygote. Consequently, MZ twins always share the same biological sex, whereas DZ twins can be either the same or different sex. Therefore, the genetic similarity within a group of SS twin pairs constitutes a weighted average of the 50% and 100% shared by DZ and MZ twins respectively. This proportion (H_{SS}) can be inferred from population twinning rates, as long as the following two assumptions are met: (1) the sample is representative of the corresponding population; and (2) the specific population's twinning rate is known and well-established.

Numerous studies have established reliable population twinning rates, which vary across countries (Monden et al., 2021; Pison et al., 2015), ethnicities (Pollard, 1995), social classes (Gómez et al., 2019; Walle et al., 1992), and eras (Esposito et al., 2022;

Gómez et al., 2019), among many other factors (Beck et al, 2021; Nylander, 1981). When both sample characteristics and the population rates are available, a weighted H_{SS} for a particular sample can be calculated to fit the ACE model. Without population rates, 'local' estimation methods, such as Weinberg's (1901) differential rule, the mixture distribution model (Neale, 2003), and latent class analysis (Heath et al., 2003), can be used to derive H_{SS} strictly from the sample attributes. Regardless of the approach taken to derive H for SS twins, the expected value of H_{SS} for the group will fall in the $0.5 < H_{SS} < 1.0$ range. Consider, for example, a univariate ACE model applied to SS and OS twins. In this case, the expected variance structure, shown in equations 1 and 2, has off-diagonal values (representing the covariance between twin pairs) in SS twins' covariance matrix for additive genetics (A) equal to H_{SS}.

$$\Sigma_{SS}(\theta) = \begin{pmatrix} 1 & H_{SS} \\ H_{SS} & 1 \end{pmatrix} A + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} C + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} E \quad (1)$$

$$\Sigma_{OS}(\theta) = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} A + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} C + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} E \quad (2)$$

By fitting the two groups of kin pairs to the (co)variance structure displayed in equations 1 and 2, we can decompose the total variance into additive genetic variance (A), common environmental variance (C), and unique environmental variance (E). The resulting variance estimates will be identical to the ACE model from the classical twin design (Rijsdijk & Sham, 2002). The respective covariance matrix of any two groups of kin pairs with different H ($\Delta H > 0$) can be used to estimate all three variance components. For example, Rodgers et al. (2019) used cousins (H = .125) and half cousins (H = .0625) from the National Longitudinal Survey of Youth to fit a series of ACE models to estimate the heritability of height. Similarly, kinship links identified in the China Family Panel Study, including twins, full siblings and cousins, also present opportunities for the application of nonclassical ACE models (Lyu & Garrison, 2022b).

Power in Designs of $\Delta H < .5$

A power analysis on SS-OS design can enhance our understanding of their statistical properties and evaluate their feasibility. In classical twin designs, the relatedness difference (ΔH) between MZ and DZ twins is .5. However, in nonclassical models, like the SS-OS design, this difference is generally less than .5. Consequently, the implied covariance matrices for these nonclassical kin groups (as represented by the 2 × 2 matrices in equations 1 and 2) tend to be more similar than in classical twin designs. The implication is that we need a narrower estimated standard error to ensure that the two covariance matrices generated from empirical data can fit the implied structure of the univariate ACE model. Put simply, researchers may need larger sample sizes to achieve the same level of statistical power as a classical twin design.

In their commentary on Scarr-Salapatek (1971), Eaves and Jinks (1972) investigated the power of standard proportion of additive genetic variance (a^2; Computed by A/(A + C + E)) estimation under a weighted least-squares approach when using SS and OS twins. Specifically, they considered the case where $a^2 = .6$ and $\Delta H = .2$, and they found that the SS-OS design needed a sample size that was approximately three times larger to achieve comparable power as the classical twin design. However, the

---

[1]To illustrate, we did a brief search on ICPSR for household surveys. Out of the 18,916 studies in ICSPR's repository, 719 fall under the 'household' subject term. Of these, a mere 111 studies incorporate keywords like 'twin' and 'zygosity', indicating that only a small fraction of household surveys include data on twins.

[2]Much has been written about violations of the equal environments assumption and its potential implications for twin studies (see Felson 2014 for a comprehensive overview and reanalysis). Such criticisms often hinge on the argument that MZ twins, due to their identical appearances and obvious 'twinness', are subject to more similar treatment, thereby inducing potential confounds in violation of the equal environments assumption. To mitigate such concerns, one could apply the SS-OS model, which effectively spreads the MZ twins across both groups.
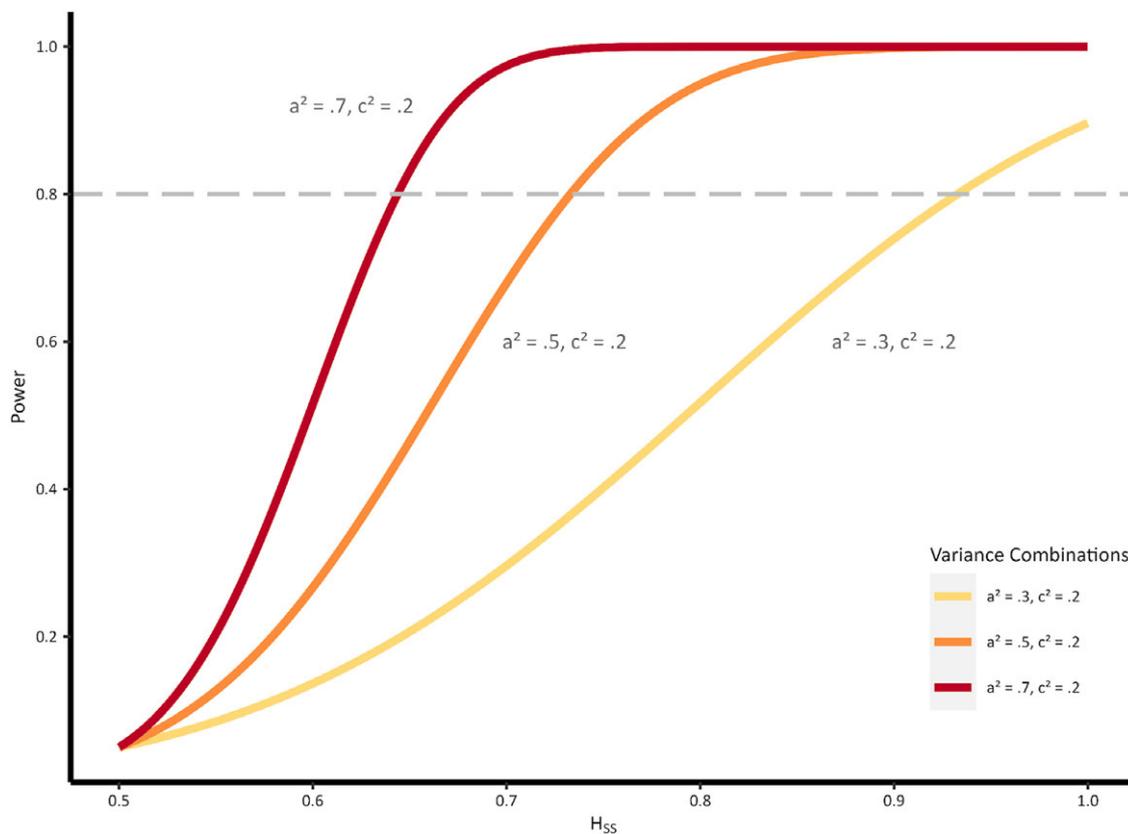
Figure 1. This figure illustrates the power for detecting a significant a^2 parameter as a function of genetic relatedness of SS twins and variance combinations based on equation 3. We have fixed the sample size to 500 and set α to .05.

statistical power of a^2 estimation is heavily associated with the variance combination and the relative proportion of MZ and DZ twins (Verhulst, 2017). As a result, the 'three times sample size' rule of thumb may not be universally applicable, and arguably should not be treated as a rule without a more systematic exploration of power in such designs.

Mathematically, we adapted Visscher's (2004) paradigm by using least-squares (LS) estimation to evaluate the power of the univariate ACE model as a function of the genetic relatedness of the SS twins (H_{SS}). Equation 3 illustrates the relation between power and sample size and R. A detailed mathematical derivation of equation 3 is provided in Appendix A.

$$\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 = n \cdot \frac{(a^2)^2(H_{SS} - 0.5)^2}{(1 - (H_{SS}a^2 + c^2)^2)^2 + (1 - (0.5a^2 + c^2)^2)^2}$$

(3)

In equation 3, a^2 is the standardized proportion of the additive genetic variance of the measured trait, c^2 is standardized proportion of the common environment variance of the trait, and n is the number of kin pairs in each kin group. $Z_{1-\alpha}$ and $Z_{1-\beta}$ denote corresponding Z values in an $N(0,1)$ for the assigned type-I error rate (α), and is power ($1-\beta$), respectively. Power is positively associated with sample size and genetic relatedness of the SS twins ($.5 < H_{SS} < 1$), provided that the SS and OS twins have the same sample size n and a specified type-I error rate.

The power for detecting additive genetic variance varies as a function of the relatedness of SS twins. Illustrated in Figure 1, when $a^2 = .3/ .5/ .7$, $c^2 = .2$, $e^2 = .5/ .3/ .1$, $n = 500$, $\alpha = .05$, power

increased as the H_{SS} twin relatedness deviated from .5. As the A component increases its share in total variance, the same level of power could be maintained when using kin pairs with smaller ΔH. Although most recent research estimates variance components with the maximum likelihood (ML) approach, the results of the univariate ACE model fit with LS and ML are very similar (Visscher, 2004). This striking similarity allows for extrapolating the association found with LS to the general pattern of univariate ACE model fitting. However, ML estimation generally has greater power than LS estimation (Visscher, 2004), leading to differences in sample size requirements for satisfactory power. Relying solely on LS analytic results does not offer researchers sufficient accuracy to establish a priori power estimation for empirical analysis. Moreover, simulations facilitate establishing different levels of c^2 to investigate their effects on the power of the a^2 estimate. Thus, we will also examine the power of the a^2 estimate in a more comprehensive set of simulations in addition to deriving power using least squares estimation.

The Challenge of Non-Positive Parameter Estimates

Besides the issue of power, using the nonclassical kin models may result in estimated variance components that are zero (Cholesky decomposition) or negative (correlated factors models; Carey, 2005). Although it is possible that these nonpositive estimates reflect biological mechanisms that arise from genetics or the environment (Steinsaltz et al., 2020), such biologically justified estimates are rare (Verhulst et al., 2019). More often, these parameters reflect something statistical — related to modeling or measurement. Negative estimates can occur due to model

misspecification. For example, misspecification between ACE and ADE models that can appear as C and D cannot be estimated simultaneously or appear when using a too simplified model (e.g., ACE) to reveal a more complex structure (Ozaki et al., 2011; Verhulst et al., 2019; see Hunter et al., 2021 for a mathematical approach discussing models with more complex structures.).

Furthermore, negative estimates can simply arise from sampling error (Tabachnick et al., 2019). This can occur when a biased observed variance or covariance in either the MZ or DZ sample reduces one of three components to zero or a negative value. Although larger samples in modern twin studies can mitigate these sampling errors, concerns persist in designs using SS and OS twins or other nonclassical designs. For example, a study examining the heritability of height used 116 SS twins and 61 OS twins from the China Family Panel Study to fit an ACE model (Lyu & Garrison, 2022b). This resulted in a slight negative E estimate. Similarly, a study using cousins (Rodgers et al., 2019) encountered several instances of zero-value estimates. In such designs, using kin pairs with genetic relatedness differences of less than .5 ($\Delta H < .5$) requires larger sample sizes to ensure that the observed covariance pattern adequately represents the population parameters. However, obtaining larger sample sizes may not always be feasible for researchers working with public datasets. At present, there is no established guidance for suitable sample sizes in an SS-OS design, leaving researchers without the ability to determine whether their specific combinations of kin groups and sample size is sufficient for their desired level of power.

Sex-Limitation Models

Another potential issue when using SS and OS twins instead of MZ and DZ twins is the effect of sex limitation (Neale & Cardon, 2013). In this context, 'sex limitation' refers to models that account for differences in genetic and environmental influences on a trait by biological sex. The difference may be scalar, indicating that all sexes are influenced by the same factor but to varying degrees, or nonscalar, indicating that specific factors influence only one of the sexes (Neale et al., 2006). Traditional twin designs often exclude OS DZ twins to avoid potential confounds introduced by sex differences (Polderman et al., 2015). However, the challenges associated with sex limitation effects become unavoidable when fitting ACE models with SS twins and OS twins. Beyond just the obvious methodological necessity, there are substantive implications. For example, past research has found that within an OS sibling pair, the male sibling often receives more parental resources than his female sibling, especially in nations with limited social resources (Blau et al., 2020; Das Gupta et al., 2003; Hesketh & Xing, 2006). In the case of OS twins, one study found that family background effects were stronger for the male twin compared to the female twin, though the genetic effects were comparable for both sexes (Miller et al., 1997). The assumption in classical twin design that the common environment (C) component is identical between twin pairs is not substantiated under these circumstances (Felson, 2014; Kendler et al., 1993; Loehlin & Nichols, 1976; Richardson & Norgate, 2005). One commonly suggested modeling solution is to set the common environment correlation at a value less than 1 (Neale et al., 2006). This adjustment aims to partially account for the impact of sex differences by implementing a sex-limited scalar in a univariate ACE model (Neale et al. 2006). Equation 4 illustrated the assumed variance structure for OS twins with sex limitations,

Table 1. Simulation of design conditions

H_{SS}*	.55, .60, .65, .70, .75, .80, .85, .90, .95, 1.00	
Sample sizes**	30, 60, 90, 150, 210, 300, 450, 750, 1200, 1950	
Variance structure	Unstandardized (A, C, E)	Standardized (a^2, c^2, e^2)
	1.5, 0.6, 0.9	50%, 20%, 30%
	2.4, 0.3, 0.3	80%, 10%, 10%
	1, 1, 1	33.3%, 33.3%, 33.3%
	0.5, 2, 0.5	16.7%, 66.6%, 16.7%

Note: *H_{SS} is the expected genetic relatedness of same-sex twins.
**Sample sizes are the number of twin pairs in each group. If the sample size is 30, there will be 30 pairs of SS twins and 30 pairs of OS twins, totaling 120 individuals.

$$\Sigma_{OS}(\theta) = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} A + \begin{pmatrix} 1 & r_c \\ r_c & 1 \end{pmatrix} C + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} E \quad (4)$$

where values of off-diagonals in the common environment (C) covariance matrix are the presumed common environment correlation (r_c). However, it is unknown how this approach will affect the power and performance of the ACE model fitting with SS twins and OS twins, or any other two groups of kin pairs where the H difference is less than .5.

Hence, the current study aims to better understand the complexities of utilizing SS twins and OS twins in genetically informed designs. Specifically, we developed a series of simulations to investigate (1) the power of heritability estimation, (2) ACE model performance in AIC-based model selection, and (3) the frequency of the negative estimates as a function of H and sample sizes under the maximum likelihood theory. In addition, we analyzed the impact of sex-limitation models within this framework.

Methods

We conducted a simulation with a $10 \times 10 \times 4$ design (see Table 1) with 1000 replications per condition. Given that our primary objective is to illustrate the impact of H_{SS} and sample size on the fitting of univariate ACE models, we have established 10 conditions for H_{SS} ranging from .55 to 1.00 in increments of .05. These 10 progressive conditions encompass the potential range of the SS-OS design. Furthermore, we have set 10 conditions for sample sizes, ranging from 30 to 1950, to cover most scenarios in empirical studies using the SS-OS design (Polderman et al. 2015). As a result, we will simulate 100 conditions varying in H_{SS} and sample sizes, providing a robust guideline for practical applications of the SS-OS design.

Previous research indicated that the power of the A estimation in a univariate ACE model highly depends on the relative scale between A and C (Verhulst, 2017). In reality, different traits have a broad range of patterns for A and C (Polderman et al., 2015). Hence, four conditions of A, C, and E variance patterns were set to cover different traits with different variance component structures. All four variance patterns have a total variance of 3. The proportion of A variance ranges from 16.7% to 80%, emulating traits subject to low, medium, and high additive genetic variance. Standardized proportions of each component in four conditions are also displayed in Table 1.

Data Generation

MZ and DZ data were simulated by generating random numbers under multivariate normal distribution by functions in ACEsimFit package 0.0.0.9 (Lyu & Garrison, 2022a). Based on the H_{SS}, a certain proportion of MZ and DZ twins were generated separately to form a group of SS twins, and then another group of DZ twins was generated as the group of OS twins. The simulated data were fitted with a univariate ACE model using the correlated factor approach and, for each condition, the simulation was repeated 1000 times. All simulations were performed in R version 4.1.3 (R Core Team, 2022). The univariate ACE models were fit using OpenMx 2.20.6 (Neale et al., 2016) with the NPSOL 5.0 optimization algorithm.

As for the investigation of modeling sex limitations, data were simulated under the same conditions, using a variance pattern of A = 1.5, C = .6, and E = .9. Notably, to emulate the sex-limited effect in the common environment, the correlation of C (r_c) between OS twins was set at .95 instead of 1.00. However, for simplicity, we did not misspecify the common environment correlation.

The framework suggested by Satorra and Saris (1985) formed the basis for deriving the power of heritability estimation. We calculated the mean noncentrality parameter (NCP), by comparing the values of log-likelihood ratio tests (-2 log likelihood) for the ACE and CE models, for the 1000 models in each condition. Next, we derived the power for each condition from a comparison between the null chi-square distribution and the alternative chi-square distribution with a given NCP. For a more detailed description of this approach, refer to Satorra and Saris (1985), Verhulst (2017), and Visscher (2004).

To evaluate how effectively the model correctly identified the assumed ACE variance structure, we employed the Akaike Information Criterion (AIC; Akaike 1998) to compare the relative performances of the ACE, AE, and CE models. We used the proportion of 1000 models where the ACE model has the lowest AIC among three models under each condition as an indicator of correct model selection.[3] Lower AIC values suggest one model's superiority in explaining the data relative to other models. AIC has been a long-used approach to evaluating the relative performance among ACE, CE, and AE models in univariate twin designs and yields adequately accurate decisions for continuous traits (Sullivan & Eaves, 2002). Furthermore, we also calculated the proportion of the 1000 models where at least one of A, C, and E estimates has a negative value to evaluate the influence of sample sizes and H on model fitting.

At the recommendation of a reviewer, we also investigated A parameter bias in the absence of H_{SS} misspecification. We computed average parameter A estimates across 1000 fitted models in every single condition under the variance combinations A = 2.4, C = .3, E = .3; A = 1.5, C = .6, E = .9 and A = 1.0, C = 1.0, E = 1.0.

Results

We ran a series of simulations to investigate the impact of H_{SS} on model performance. We summarized the simulation results of the power of heritability estimation, ACE model performance in AIC Based Model Selection and the frequency of negative estimates of the 1000 fitted models in each condition. We presented the results

---

[3]Ideally, since the simulated data are generated based on the assumption that A, C, and E all contribute to the outcome score, the variance structure should be best explained by the ACE model.

in a series of matrices where the x-axis displays the 10 conditions of sample sizes and the y-axis is 10 conditions of H_{SS}. Interpretation and insights from the results were discussed for the three criteria separately. Because many of the result patterns were similar, we primarily presented the results of A = 2.4, C = .3, E = .3 (80%, 10%, 10%) and A = 1.5, C = .6, E = .9 (50%, 20%, 30%). The results and corresponding figures for the other two combinations are available in Appendix B.

Power of Heritability Estimation

Generally, we found that the power of a univariate ACE model to detect A is positively associated with sample size and H_{SS}. This finding was consistent with our mathematical derivation from the LS approach. As shown in both Figure 2-1 and Figure 2-2, the positive association between power and H_{SS} suggests that a higher proportion of MZ twins in the SS twins will require a smaller sample size to reach a power of .8. For example, when the variance combination is A = 1.5, C = .6, E = .9 (Figure 2-1), a sample with H_{SS} = .75 needs about 450 pairs of SS and OS twins to reach a power of .8, whereas a sample with H_{SS} = .90 only needs 150 pairs to reach .8. As the covariance structures of SS and OS twins become more dissimilar, smaller samples will be sufficient to distinguish the covariance structures. Conversely, as they grow more similar, a larger sample is needed to have the same effect. Although the positive association is similar for two combinations of variance components, each condition for the combination of A = 2.4, C = .3, E = .3 (Figure 2-2) demonstrated higher power compared to the corresponding condition for the combination of A = 1.5, C = .6, E = .9 (Figure 2-1). For example, in a condition H_{SS} = .75 and N = 300, the power for the variance combination of A = 2.4, C = .3, E = .3 is .734, which is lower than a power of .997 for the variance structure of A = 2.4, C = .3, E = .3 in the same condition. The proportions can be interpreted as we have power of .734 to detect a significant difference between the estimated value of A and 0 for the variance combination of A = 2.4, C = .3, E = .3. In other words, out of the 1000 models, 734 of them found the expected significant effect. Comparing all four variance combinations, we find that a greater share of A in total variance is associated with higher power in each condition. This finding is consistent with our mathematical derivation and Verhulst's (2017) results that the power of the ACE model is higher when both the proportion of A and C in the total variance increase.

Model Performance: AIC-Based Model Selection

Regarding model performance in AIC-based model selection, we found that H_{SS} and sample sizes generally but not exclusively have a positive association with the overall model performance. Model performance was operationalized by counting the percentage of the 1000 models where the ACE models have lower AIC values compared to AE and CE models. For example, with the variance combination of A = 1.5, C = .6, E = .9 (Figure 3), a sample with H_{SS} = .75 needs about 1200 pairs of SS and OS twins to reach a power of .8, whereas an H_{SS} = .90 sample only needs 450 pairs to reach .8. More specifically, in both variance combinations of A = 1.5, C = .6, E=.9 (Figure 3) and A = 2.4, C = .3, E = .3 (Supplementary Figure S2-1 in Appendix B), the worst conditions occur in the middle of the grid, where neither the sample size nor H_{SS} are extremely small. Although the conditions on the upper left of the grid are better than the middle ones, the overall power at that range is far from acceptable. For all the variance combinations,
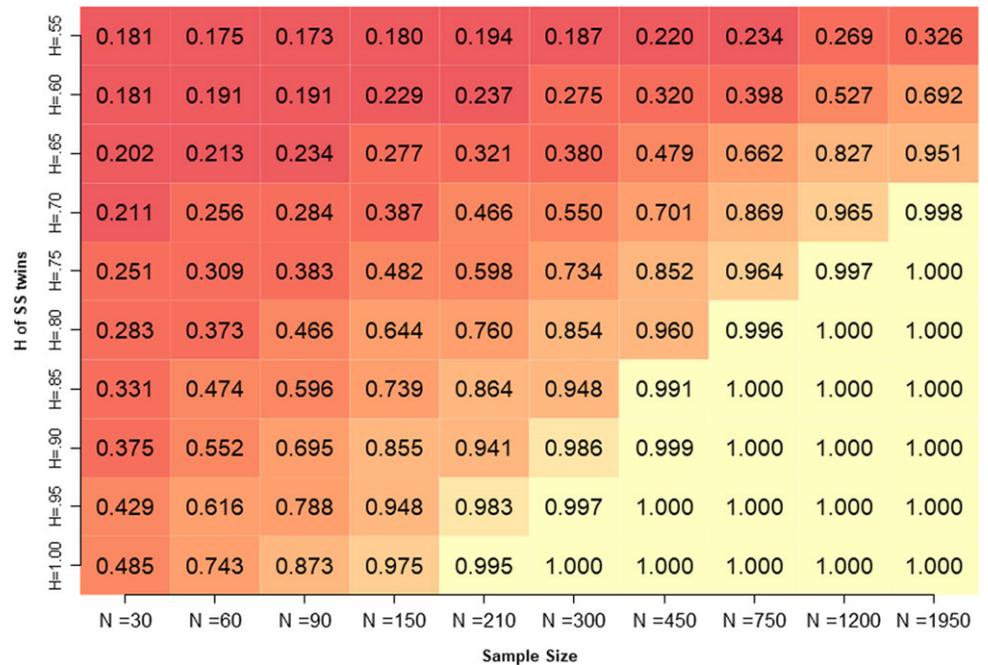
Figure 2-1. Illustrated here is the power of the ACE model to detect A under the simulated variance of A = 1.5, C = .6, E = .9 (50%, 20%, 30% respectively), as a function of sample size per twin group and H of SS twins. Power in each cell was calculated based on the average non-centrality parameter of 1000 simulations under the corresponding condition. Darker cell colors denote lower power.
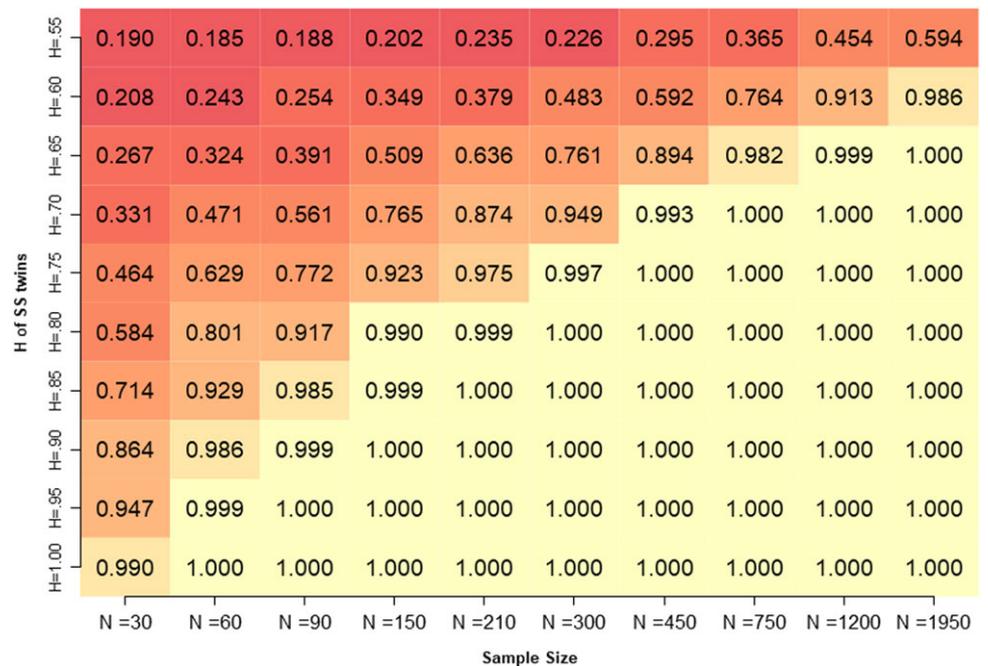
Figure 2-2. Illustrated here is the power of the ACE model to detect A under the simulated variance of A = 2.4, C = .3, E = .3 as a function of sample size per twin group and H of SS twins. Power in each cell was calculated based on the average noncentrality parameter of 1000 simulations under the corresponding condition. Darker cell colors denote lower power. 'Sample size' indicates the number of kin pairs in each kin group.

acceptable overall power only exists when sample sizes and H_{SS} were relatively large, which is consistent with our prediction.

We noticed that the association pattern between H_{SS} and sample sizes and model performance fluctuated across the different variance component combinations. A relatively vague trend shows that as the C variance component dwindles, the model performance in each condition deteriorates. One intuitive explanation is that when the share of C decreases, more information (larger sample size) is needed to distinguish the covariance structure of the ACE model from the AE model. Consequently, the variance combination of A = .5, C = 2.0, E = .5 (Supplementary Figure S2-3 in Appendix B) had the best model performance results among the four combinations. An alternative possible explanation might be that an increase in E leads to a decline in model performance. Another interesting pattern emerges when the share of the C component increases: the 'gorge' in the fittings results moves towards the upper left, along with improved model performance.

Frequency of Negative Estimates

Our results indicate that negative estimates for A, C or E are less frequent with increasing H_{SS} and sample sizes. For example, given a variance combination of A = 1.5, C = .6, E = .9 (Figure 4), a sample with H_{SS} = .75 needs about 300 pairs of SS and another 300 pairs of OS twins to reduce the frequency

Figure 3. Illustrated here is the proportion of the fitting results from 1000 simulated datasets where the ACE model has the lowest AIC compared to the AE and CE models. Simulated variance was set at A = 1.5, C = .6, E = .9 (50%, 20%, 30% respectively), as a function of sample size per twin group and H of SS twins. Darker cell colors denote lower power. 'Sample size' indicates the number of kin pairs in each kin group.

Figure 4. Illustrated here is the proportion of fitting results from 1000 simulated datasets with at least one negative estimate for A, C or E variance components, when variance is set to A = 1.5, C = .6, E = .9 (50%, 20%, 30% respectively), as a function of sample size per twin group and H of SS twins. Darker cell colors indicate higher prevalence of negative estimates. 'Sample size' indicates the number of kin pairs in each kin group.

of negative estimates to a 10% level. In contrast, a sample with H_{SS} = .90 only needs 150 pairs. The 10% frequency indicates that out of the 1000 simulated models at least one negative parameter estimate occurs in 10% of the models. It appears that larger sample sizes and higher H_{SS} values reduce the likelihood of negative estimates due to sampling error. Additionally, the negative estimates seem to be rather sensitive to different combinations of variance components. For example, there are distinctly fewer negative estimates for a variance combination of A = 1.5, C = .6, E = .9 (Figure 4) than for A = 2.4, C = .3, E = .3 (Figure S3-1). For example, under the conditions H_{SS} = .75 and N = 300, the frequency of negative estimates is 10.1% for the

variance combination of A = 1.5, C = .6, E = .9, as compared to 23.6% for the variance structure of A = 2.4, C = .3, E = .3. Given the smaller proportion of C and E components in the total variance for the latter combination, the chance of obtaining a negative variance estimate due to sampling error increases compared to A = 1.5, C = .6, E = .9.

Sex-Limited Effects (r_c = .95; A = 1.5; C = .6; E = .9)

Addressing the potential for sex-limited effects, our results suggested that the general positive association between H_{SS} and sample sizes and three criteria of model performance was

Figure 5. Displayed here is the power of the ACE model to detect A under the simulated variance of A = 1.5, C = .6, E = .9 (50%, 20%, 30% respectively) and the sex-limitation scalar of r_c = .95 included as a function of sample size per twin group and H of SS twins. Power in each cell was calculated based on the average noncentrality parameter of 1000 simulations under the corresponding condition. Darker cell colors denote lower power. 'Sample size' indicates the number of kin pairs in each kin group.

broadly consistent with the standard model without sex-limited effects. The power to detect A diminished slightly when we set the correlation between OS twins to .95 (Figure 5), compared to the results with the same variance components (A = 1.5, C = .6, E = .9) without considering sex-limitation (Figure 2-1). A decrease in r_c corresponded with a reduction in the proportion of C in the total variance. In turn, to achieve the same power level, larger sample sizes are required, which is consistent with Verhulst's (2017) findings. Additionally, the models that factored in sex-limited effects (Figure 5) yielded more negative estimates than in the standard models (Supplementary Figure S4-3). A comparison of Figure 5 and Supplementary Figure S4-2 revealed an interesting pattern in the overall model fitting. When the sample sizes are relatively small, models incorporating sex-limited effects suggested a worse overall fit compared to models that did not. However, when sample sizes exceeded 450 pairs per group, the models with sex-limitation outperformed the standard models.

Parameter Bias

Following a reviewer's suggestion, we investigated the bias of the 'A' parameter, computing average 'A' estimates across 1000 simulated models for each condition. Because the pattern of results was similar across conditions, we present one condition in Figure 6, with others in the Supplementary Appendix B. This figure depicts the 'A' parameter bias under the variance distribution A = 1.5, C = .6, E = .9 (50%, 20%, 30% respectively), showing that A estimates are not biased from 1.5 drastically in all conditions. A estimates are slightly biased (i.e., A estimates deviate from 1.5 upwards or downwards by 1% of the total variance, which equates to .03 in our study) when H_{SS} is small or when sample sizes are restricted. More specifically, A is inclined to be underestimated when H_{SS} is below .65 and the sample size falls short of 300, as illustrated in the upper-left part of Figure 6. In contrast, A tends to be overestimated when H_{SS} exceeds .65 but the sample size is less than 210 (lower-left part of Figure 6),

or when H_{SS} is below .65, but the sample size exceeds 300 (upper-right part of Figure 6). As expected, the estimation bias for the A parameter gradually diminishes with higher H_{SS} values and larger sample sizes. In general, a sample size above 300 and an H_{SS} value greater than .65 can help to avoid the presence of unbiased estimates in the lower-right triangle of Figure 6.

Discussion

In the current study, we investigated how well univariate ACE models perform to correctly estimate the variance structure of A, C and E as a function of the expected relatedness of the SS twins (H_{SS}) and sample size. We adopted Visscher's (2004) LS paradigm to mathematically derive the positive relationship among power, H_{SS}, and sample sizes. We conducted simulations to further explore how heritability power, AIC-based model performance, and reduction of negative estimates are positively associated with larger H_{SS} and larger sample sizes. In addition, we examined whether the simple solution of changing the common environment correlation to .95 for addressing sex-limited effects impacted model performance. We found that the solution causes slightly worse model performance under most circumstances.

Both the algebraic derivations and simulations illustrated a positive relationship between H_{SS} and the power of correctly detecting the additive genetic effects (A) in an ACE model. A larger difference between the genetic correlations (ΔH) will require less information to distinguish the covariance structure of the SS twins from the covariance structure of the OS twins, as the only difference in the implied covariance structure between SS twins and OS twins is the correlation for additive genetics. The difference can also be understood as the significant difference between a model where additive genetics plays a role in affecting the phenotype from a model where additive genetics does not. We also found that traits subject to more additive genetic influence would have higher power under all conditions of H_{SS} and sample sizes. Our results are consistent with previous findings (Verhulst, 2017). Mathematically, an increase of standardized additive genetic

Figure 6. Average estimates of 'A' obtained from 1000 models, each fit to simulate data with variance combination A = 1.5, C = .6, E = .9 (50%, 20%, 30%). Darker cell colors denote larger deviations from the population parameter A = 1.5. 'Sample size' indicates the number of kin pairs in each kin group.

component (approximately a decrease in the proportion of error variance) would lead to a greater difference between the intraclass correlations for OS and SS twins, which would eventually contribute to a higher power (see a more detailed math derivation in Visscher 2004).

We found a similar positive association between H_{SS} and AIC-based model selection. Model performance, distinct from the power to detect the A parameter, evaluates whether the ACE model has the lowest AIC value. Typically, a high AIC-based model performance requires a proper fit of A, C and E components concurrently. This model performance offers a more conservative model selection criteria than a single estimate's power. The variance components' structure also influences the model performance. We observed that a higher proportion of C in the total variance suggested higher power across all conditions. Our results suggested that an adequate amount of C is vital for the model to correctly distinguish between ACE and AE models, because in our results the correct models (ACE) were more often misspecified as AE models than AC models. Nevertheless, further algebraic and simulation research is needed to identify factors impacting AIC comparison approaches.

For negative estimates, our study demonstrated that in general when the relatedness difference between two modeled groups (ΔH) is less than .5, negative estimated parameters are not unusual, even when samples were relatively large. Although the conventional wisdom is that the estimated error variance should always be non-negative, that reasoning is based on the idea that within-pair variance can never be eliminated. Our study highlighted that negative E estimates can occur simply due to sampling errors in some special circumstances. For example, suppose we fit an ACE model with a small number of kin pairs to the target trait predominantly affected by genes and shared environment. In that case, the E parameter will have a wide confidence interval. Therefore, it is not unusual that the model will estimate a negative E. Although we could force the estimate to be non-zero, that results in more problems. Indeed, ACE models with explicit or implicit constraints on estimates can cause deviations from the

assumed type-I error rates and lead to biased estimates (Verhulst et al., 2019). Therefore, we do not recommend forcing the negative estimates to be greater than zero, especially in circumstances where they are not unusual. In our study, the negative estimates were entirely the result of sampling error, and occurred when variance components were relatively close to zero. Under those circumstances, estimates are more likely to be negative. As a corollary, in empirical studies, encountering negative estimates is not synonymous with a failed model. Rather, negative estimates can be an indicator of low power, small effect sizes, or general model misspecification. Therefore, we recommend checking other criteria given the specific conditions before continuing to analyze the results, adjusting model specifications, or discarding the data entirely.

We found that A estimates were slightly biased when the sample sizes were small or ΔH was low. Much like other analyses, greater ΔH and larger sample sizes contribute to reduced bias of A estimates, reaffirming the ideal situation for the SS-OS design: a sample size exceeding 300 pairs per group. Further, we found no systematic bias in this design, meaning that any biased conditions are likely the result of randomness in the simulation and model-fitting processes. An interesting future direction to explore is the sensitivity of this design to H_{SS} misspecification. Given that H_{SS} is usually an estimated value derived from population twinning rates or local estimating algorithms rather than a population parameter, we suspect that various degrees of H_{SS} misspecification could substantially affect parameter bias.

Although our study focused primarily on the SS-OS design, these results are applicable to other research designs where the difference in relatedness (ΔH) is less than .5. Another scenario where H can diverge from .5 arises when researchers intend to fit covariance structure models, like the ACE model, with nontwin kin pairs. Such datasets can also support fitting an ACE model with MZ twins, siblings, or distant cousins. These configurations also result in an H difference not equal to .5. For example, past studies have employed full siblings and cousins to estimate heritability for specific phenotypic outcomes (Chakraborty et al., 1977; Rodgers et al., 2019; Souto et al., 2000). The

difference in H between cousins, siblings or twins is invariably less than .5, given that the relatedness coefficient for cousins does not exceed .125. These nontwin designs can serve as a valuable resource for researchers investigating the environmental and genetic influence on various traits.

Future researchers planning to use two groups of kin pairs with a ΔH less than .5 should at a minimum avoid scenarios with a ΔH less than .1 and sample sizes smaller than 60 pairs per group. Since we found the association between model performance and H and sample sizes varied a lot along with the variance component structure of the targeted trait, proposing a single guideline for all circumstances will be inappropriate. Indeed, numerous studies have warned against overreliance on rules of thumb in structural equation models (Chen et al., 2008; Heene et al., 2011; Kyriazos, 2018; Montoya & Edwards, 2021), including within behavior genetics (Garrison & Rodgers, 2021). Instead, using available parameters to calculate the power of the heritability estimation before using empirical data to fit the ACE model will be preferable. If one study does not have a specific focus on A or C but is designed to illustrate the multiple sources of effects, an overall model-fit indicator like AIC in our study would be a more appropriate reference. Nevertheless, as the criteria like AIC could only be evaluated using simulations, researchers could look up the supplementary tables mentioned in the Supplementary Appendix B to find an approximate power rate corresponding to the parameter setting in their own study. Alternatively, we encourage researchers to run their own simulations using the expected parameters and covariance structure. That simulation will lead to a tailored recommendation indicating what proportion of the nested comparisons suggest the ACE structure is the best-fit model. We developed the ACEsimFit package to assist such encouraged researchers. It contains several R functions and vignettes demonstrating how to simulate and fit the models (Lyu & Garrison, 2022a).

Our results indicated less robust models when addressing sex-limited effects by slightly decreasing the assumed common environmental correlation between OS twins. However, sex-limited effects are far more complicated than a reduction of common environmental correlations. For example, in a study using SS and OS twins, OS twins may not have exactly the same family environment as the classical twin study assumed due to gender inequality (Blau et al., 2020; Das Gupta et al., 2003; Hesketh & Xing, 2006). From a modeling perspective, both genetic and environmental differences between sexes can take different forms, such as scalar and nonscalar sex limitations (Neale et al., 2006). From an empirical perspective, different traits may be susceptible to sex-limited effects. For example, height and BMI are traits that have substantial sex differences in their heritability (Hesketh & Xing, 2006; Schousboe et al., 2003; Silventoinen et al., 2001) but personality traits such as the Big 5 do not (South et al., 2018). Hence, before using the SS and OS twins to fit univariate ACE models, we recommend carefully considering the specific potential impacts of sex-limited effects. We also recommend addressing them by either modifying the assumed component structure or considering alternative models (e.g., a G × E model or a model assigning different covariance structures by biological sex; Neale et al., 2006).

Our study assessed the feasibility and risks of using twin pairs with smaller genetic relatedness differences in univariate ACE models. However, like all simulations, we had to keep the simulation scope narrow. First, we only evaluated univariate ACE models. Some research questions can only be addressed with the multivariate models, such as examining covariance between multiple traits and estimating A, C, D and E simultaneously (Maes et al., 2021). The increased complexity of multivariate models likely demands larger sample sizes or ΔH for comparable power, but further investigation is needed. Second, our derivations and simulations assume that H_{SS} is not misspecified and that the observed phenotype is normally distributed, conditions that may not always be met in empirical settings. Approaches such as population twinning rates, Weinberg's differential rule (Weinberg, 1901), mixture distribution models (Neale, 2003), and latent class analysis (Heath et al., 2003) give an approximation, not a direct observation, of MZ twins' proportion in SS twins, potentially biasing the H_{SS}. Previous research has suggested that the misspecification of H_{SS} and non-normal distributions could bias estimated parameters (Benyamin et al., 2006), indicating a potential avenue for future research. Therefore, future efforts should be made to investigate the impact of parameter misspecification and non-normal distributions on the associations between H_{SS} and model performance. Third, although AIC has been widely used in behavior genetics to determine the 'best model' (Sullivan & Eaves, 2002), its accuracy as a selection approach remains under-examined. A lengthy appendix in Garrison and Rodgers (2021) hints at potential issues with AIC as a selection criterion, and the worst-fitting 'gorge' seen across all AIC result matrices further point to potential shortcomings of this approach. Therefore, more comprehensive research should be done to investigate this model selection approach.

Conclusion

In the current study, we have identified several factors that impact the performance of the ACE model. To begin with, we found that the power to detect a significant additive genetic (A) component was positively associated with the difference in genetic relatedness of two kin groups (ΔH) and sample size. Similarly, we noted a positive association between the ACE model's performance — evaluated using the Akaike Information Criterion (AIC) and the lower frequency of negative estimates of ACE variance components — and both the difference in genetic relatedness (ΔH) between two kin groups and the sample size.

We observed that while different combinations of A, C and E variance followed a similar overall pattern — in that, for instance, a higher A parameter would consistently exhibit higher power at larger sample sizes — the absolute performance varied considerably. We also found that factoring in sex differences by reducing the assumed correlation of the common environment to .95 resulted in a model performance slightly inferior to the raw ACE model. Researchers using kin groups with ΔH of less than .5 should carefully consider the performance implications for their specific ACE model. It is crucial to conduct a comprehensive power analysis before delving into the interpretation of model outcomes.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Beck, J. J., Bruins, S., Mbarek, H., Davies, G. E., & Boomsma, D. I. (2021). Biology and genetics of dizygotic and monozygotic twinning. in A. Khalil, L. Lewi, & E. Lopriore (Eds.), *Twin and higher-order pregnancies* (pp. 31–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-47652-6_3

Benyamin, B., Deary, I. J., & Visscher, P. M. (2006). Precision and bias of a normal finite mixture distribution model to analyze twin data when zygosity is unknown: Simulations and application to IQ phenotypes on a large sample of twin pairs. *Behavior Genetics*, 36, 935–946. https://doi.org/10.1007/s10519-006-9086-3

Blau, F. D., Kahn, L. M., Brummund, P., Cook, J., & Larson-Koester, M. (2020). Is there still son preference in the United States? *Journal of Population Economics*, 33, 709–750. https://doi.org/10.1007/s00148-019-00760-7

Carey, G. (2005). Cholesky problems. *Behavior Genetics*, 35, 653–665. https://doi.org/10.1007/s10519-005-5355-9

Chakraborty, R., Schull, W. J., Harburg, E., Schork, M. A., & Roeper, P. (1977). Heredity, stress and blood pressure, a family set method — V: Heritability estimates. *Journal of Chronic Diseases*, 30, 683–699. https://doi.org/10.1016/0021-9681(77)90025-X

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36, 462–494. https://doi.org/10.1177/0049124108314720

Chow, S.-C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315183084

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Das Gupta, M., Zhenghua, J., Bohua, L., Zhenming, X., Chung, W., & Hwa-Ok, B. (2003). Why is Son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea. *The Journal of Development Studies*, 40, 153–187. https://doi.org/10.1080/00220380412331293807

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, 86, 130–147. https://doi.org/10.1037/0022-3514.86.1.130

Descôteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology*, 3, 28–34. https://doi.org/10.20982/tqmp.03.2.p028

Eaves, L. J., & Jinks, J. L. (1972). Insignificance of evidence for differences in heritability of IQ between races and social classes. *Nature*, 240, Article 5376. https://doi.org/10.1038/240084a0

Esposito, G., Dalmartello, M., Franchi, M., Mauri, P. A., Cipriani, S., Corrao, G., & Parazzini, F. (2022). Trends in dizygotic and monozygotic spontaneous twin births during the period 2007-2017 in Lombardy, Northern Italy: A population-based study. *Twin Research and Human Genetics*, 25, 149–155. https://doi.org/10.1017/thg.2022.19

Felson, J. (2014). What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption. *Social Science Research*, 43, 184–199. https://doi.org/10.1016/j.ssresearch.2013.10.004

Figlio, D., Guryan, J., Karbownik, K., & Roth, J. (2014). The effects of poor neonatal health on children's cognitive development. *American Economic Review*, 104, 3921–3955. https://doi.org/10.1257/aer.104.12.3921

Figlio, D. N., Freese, J., Karbownik, K., & Roth, J. (2017). Socioeconomic status and genetic influences on cognitive development. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 13441–13446. https://doi.org/10.1073/pnas.1708491114

Garrison, S. M., & Rodgers, J. L. (2021). *Fitting problems: Evaluating model fit in behavior genetic model.* https://doi.org/10.31234/osf.io/qys83

Gómez, N., Sosa, A., Corte, S., & Otta, E. (2019). Twinning rates in Uruguay between 1999 and 2015: Association with socioeconomic and demographic factors. *Twin Research and Human Genetics*, 22, 56–61. https://doi.org/10.1017/thg.2018.70

Hagenbeek, F. A., Hirzinger, J. S., Breunig, S., Bruins, S., Kuznetsov, D. V., Schut, K., Odintsova, V. V., & Boomsma, D. I. (2023). Maximizing the value of twin studies in health and behaviour. *Nature Human Behaviour*, 7, Article 6. https://doi.org/10.1038/s41562-023-01609-6

Heath, A. C., Nyholt, D. R., Neuman, R., Madden, P. A. F., Bucholz, K. K., Todd, R. D., Nelson, E. C., Montgomery, G. W., & Martin, N. G. (2003). Zygosity diagnosis in the absence of genotypic data: An approach using latent class analysis. *Twin Research and Human Genetics*, 6, 22–26. https://doi.org/10.1375/twin.6.1.22

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16, 319–336. https://doi.org/10.1037/a0024917

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. https://doi.org/10.1017/S0140525X0999152X

Hesketh, T., & Xing, Z. W. (2006). Abnormal sex ratios in human populations: Causes and consequences. *Proceedings of the National Academy of Sciences*, 103, 13271–13275. https://doi.org/10.1073/pnas.0602203103

Holden, L. R., Haughbrook, R., & Hart, S. A. (2022). Developmental behavioral genetics research on school achievement is missing vulnerable children, to our detriment. *New Directions for Child and Adolescent Development*, 2022, 47–55. https://doi.org/10.1002/cad.20485

Hunter, M. D., Garrison, S. M., Burt, S. A., & Rodgers, J. L. (2021). The analytic identification of variance component models common to behavior genetics. *Behavior Genetics*, 51, 425–437. https://doi.org/10.1007/s10519-021-10055-x

Jackson, D. L., Gillaspy Jr., J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14, 6–23. https://doi.org/10.1037/a0014694

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness. *Behavior Genetics*, 23, 21–27. https://doi.org/10.1007/BF01067551

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, 09, Article 08. https://doi.org/10.4236/psych.2018.98126

Levine, M., & Ensom, M. H. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, 21, 405–409. https://doi.org/10.1592/phco.21.5.405.34503

Loehlin, J. C., & Nichols, R. C. (1976). Heredity, environment, and personality: A study of 850 sets of twins. In *Heredity, Environment, and Personality*. University of Texas Press. https://doi.org/10.7560/730038

Lyu, X., & Garrison, S. M. (2022a). *ACEsimFit: ACE Kin Pair Data Simulations and Model Fitting* (0.0.0.9) [Computer software]. https://cloud.r-project.org/web/packages/ACEsimFit/index.html

Lyu, X., & Garrison, S. M. (2022b). Leveraging the China Family Panel Study: An estimation of height using preliminary kinship links. *Behavior Genetics*, 52, 375–375. https://doi.org/10.1007/s10519-022-10119-6

Maes, H. H., Neale, M. C., Kirkpatrick, R. M., & Kendler, K. S. (2021). Using multimodel inference/model averaging to model causes of covariation between variables in twins. *Behavior Genetics*, 51, 82–96. https://doi.org/10.1007/s10519-020-10026-8

Martin, N. G., Eaves, L. J., Kearsey, M. J., & Davies, P. (1978). The power of the classical twin study. *Heredity*, 40, Article 1. https://doi.org/10.1038/hdy.1978.10

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Milhollen, M., Lyu, X., & Garrison, S. M. (2022). The China Family Panel Study: An opportunity to combat WEIRDNESS in behavior genetics. *Behavior Genetics*, 52, 378–379. https://doi.org/10.1007/s10519-022-10119-6

Miller, P., Mulvey, C., & Martin, N. (1997). Family characteristics and the returns to schooling: Evidence on gender differences from a sample of Australian twins. *Economica*, *64*, 119–136. https://doi.org/10.1111/1468-0335.00067

Monden, C., Pison, G., & Smits, J. (2021). Twin Peaks: More twinning in humans than ever before. *Human Reproduction*, *36*, 1666–1673. https://doi.org/10.1093/humrep/deab029

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*, 413–440. https://doi.org/10.1177/0013164420942899

Neale, M. C. (2003). A finite mixture distribution model for data collected from twins. *Twin Research*, *6*, 235–239. https://doi.org/10.1375/136905203765693898

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549. https://doi.org/10.1007/s11336-014-9435-8

Neale, M. C., Røysamb, E., & Jacobson, K. (2006). Multivariate genetic analysis of sex limitation and G × E interaction. *Twin Research and Human Genetics*, *9*, 481–489. https://doi.org/10.1375/twin.9.4.481

Neale, M., & Cardon, L. R. (2013). *Methodology for genetic studies of twins and families*. Springer Science & Business Media.

Nylander, P. P. S. (1981). The factors that influence twinning rates. *Acta Geneticae Medicae et Gemellologiae*, *30*, 189–202. https://doi.org/10.1017/S0001566000007650

Ozaki, K., Toyoda, H., Iwama, N., Kubo, S., & Ando, J. (2011). Using non-normal SEM to resolve the ACDE model in the classical twin design. *Behavior Genetics*, *41*, 329–339. https://doi.org/10.1007/s10519-010-9386-5

Parsaeian, M., Mahdavi, M., Saadati, M., Mehdipour, P., Sheidaei, A., Khatibzadeh, S., Farzadfar, F., & Shahraz, S. (2021). Introducing an efficient sampling method for national surveys with limited sample sizes: Application to a national study to determine quality and cost of healthcare. *BMC Public Health*, *21*, 1414. https://doi.org/10.1186/s12889-021-11441-0

Pison, G., Monden, C., & Smits, J. (2015). Twinning rates in developed countries: Trends and explanations. *Population and Development Review*, *41*, 629–649. https://doi.org/10.1111/j.1728-4457.2015.00088.x

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*, Article 7. https://doi.org/10.1038/ng.3285

Pollard, R. (1995). Ethnic comparison of twinning rates in California. *Human Biology*, *67*, 921–931.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*, Article 7624. https://doi.org/10.1038/538161a

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richardson, K., & Norgate, S. (2005). The equal environments assumption of classical twin studies may not hold. *British Journal of Educational Psychology*, *75*, 339–350. https://doi.org/10.1348/000709904X24690

Rijsdijk, F. V., & Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in Bioinformatics*, *3*, 119–133. https://doi.org/10.1093/bib/3.2.119

Rodgers, J. L., Beasley, W. H., Bard, D. E., Meredith, K. M., D Hunter, M., Johnson, A. B., Buster, M., Li, C., May, K. O., Garrison, S. M., Miller, W. B., van den Oord, E., & Rowe, D. C. (2016). The NLSY kinship links: Using the NLSY79 and NLSY-children data to conduct genetically-informed and family-oriented research. *Behavior Genetics*, *46*, 538–551. https://doi.org/10.1007/s10519-016-9785-3

Rodgers, J. L., Garrison, S. M., O'Keefe, P., Bard, D. E., Hunter, M. D., Beasley, W. H., & van den Oord, E. J. C. G. (2019). Responding to a 100-year-old challenge from Fisher: A biometrical analysis of adult height in the NLSY data using only cousin pairs. *Behavior Genetics*, *49*, 444–454. https://doi.org/10.1007/s10519-019-09967-6

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90. https://doi.org/10.1007/BF02294150

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science*, *174*, 1285–1295.

Schousboe, K., Willemsen, G., Kyvik, K. O., Mortensen, J., Boomsma, D. I., Cornes, B. K., Davis, C. J., Fagnani, C., Hjelmborg, J., Kaprio, J., Lange, M. de, Luciano, M., Martin, N. G., Pedersen, N., Pietiläinen, K. H., Rissanen, A., Saarni, S., Sørensen, T. I. A., van Baal, G. C. M., & Harris, J. R. (2003). Sex differences in heritability of BMI: A comparative study of results from twin studies in eight countries. *Twin Research and Human Genetics*, *6*, 409–421. https://doi.org/10.1375/twin.6.5.409

Sham, P. C., Purcell, S. M., Cherny, S. S., Neale, M. C., & Neale, B. M. (2020). Statistical power and the classical twin design. *Twin Research and Human Genetics*, *23*, 87–89. https://doi.org/10.1017/thg.2020.46

Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J., & Rose, R. J. (2001). Sex differences in genetic and environmental factors contributing to body-height. *Twin Research and Human Genetics*, *4*, 25–29. https://doi.org/10.1375/twin.4.1.25

South, S. C., Jarnecke, A. M., & Vize, C. E. (2018). Sex differences in the Big Five model personality traits: A behavior genetics exploration. *Journal of Research in Personality*, *74*, 158–165. https://doi.org/10.1016/j.jrp.2018.03.002

Souto, J. C., Almasy, L., Borrell, M., Garí, M., Martínez, E., Mateo, J., Stone, W. H., Blangero, J., & Fontcuberta, J. (2000). Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation*, *101*, 1546–1551. https://doi.org/10.1161/01.cir.101.13.1546

Steinsaltz, D., Dahl, A., & Wachter, K. W. (2020). On negative heritability and negative estimates of heritability. *Genetics*, *215*, 343–357. https://doi.org/10.1534/genetics.120.303161

Sullivan, P. F., & Eaves, L. J. (2002). Evaluation of analyses of univariate discrete twin data. *Behavior Genetics*, *32*, 221–227. https://doi.10.1023/a:1016025229858.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (7th ed.). Pearson.

United Nations. (2008). *Designing household survey samples: Practical guidelines*. https://doi.org/10.18356/f7348051-en

Verhulst, B. (2017). A power calculator for the classical twin design. *Behavior Genetics*, *47*, 255–261. https://doi.org/10.1007/s10519-016-9828-9

Verhulst, B., Prom-Wormley, E., Keller, M., Medland, S., & Neale, M. C. (2019). Type I error rates and parameter bias in multivariate behavioral genetic models. *Behavior Genetics*, *49*, 99–111. https://doi.org/10.1007/s10519-018-9942-y

Visscher, P. M. (2004). Power of the classical twin design revisited. *Twin Research*, *7*, 505–512. https://doi.org/10.1375/1369052042335250

Visscher, P. M., Gordon, S., & Neale, M. C. (2008). Power of the classical twin design revisited: II Detection of common environmental variance. *Twin Research and Human Genetics*, *11*, 48–54. https://doi.org/10.1375/twin.11.1.48

Walle, E. V. de, Pison, G., & Sala-Diakanda, M. (1992). *Mortality and society in Sub-Saharan Africa*. https://books.google.com/books/about/Mortality_and_Society_in_Sub_Saharan_Afr.html?id=0ni3AAAAIAAJ

Weinberg, W. (1901). Beiträge zur Physiologie und Pathologie der Mehrlingsgeburten beim Menschen. *Archiv für die gesamte Physiologie des Menschen und der Tiere*, *88*, 346–430. https://doi.org/10.1007/BF01657695

Xie, Y., & Hu, J. (2014). An introduction to the China Family Panel Studies (CFPS). *Chinese Sociological Review*, *47*, 3–29. https://doi.org/10.2753/CSA2162-0555470101.2014.11082908